

APPLICATION FOR UNITED STATES LETTERS PATENT

For

**SYSTEM AND METHOD TO SUPPORT PLATFORM FIRMWARE AS A
TRUSTED PROCESS**

Inventors:

Vincent J. Zimmer
Willard M. Wiseman
Jing Li

Prepared by:

BLAKELY SOKOLOFF TAYLOR & ZAFMAN LLP
12400 Wilshire Boulevard
Los Angeles, CA 90025-1026
(206) 292-8600

Attorney's Docket No.: 42P18501

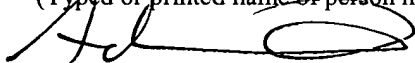
"Express Mail" mailing label number: EV320119373US

Date of Deposit: February 25, 2004

I hereby certify that I am causing this paper or fee to be deposited with the United States Postal Service
"Express Mail Post Office to Addressee" service on the date indicated above and that this paper or fee has been
addressed to Mail Stop Patent Application, Commissioner for Patents, Washington, D. C. 20231

Adrian Villarreal

(Typed or printed name of person mailing paper or fee)



(Signature of person mailing paper or fee)

February 25, 2004

(DATE SIGNED)

SYSTEM AND METHOD TO SUPPORT PLATFORM FIRMWARE AS A TRUSTED PROCESS

FIELD OF THE INVENTION

[0001] The field of invention relates generally to computer systems and, more specifically but not exclusively relates to techniques for supporting execution of platform firmware as a trusted process.

BACKGROUND INFORMATION

[0002] The past few years have seen an ever-increasing level of attacks on computer systems and servers. Malicious hackers spend hours on end trying to identify security holes via which they can embed viruses, Trojans, *etc.* Almost as soon as an operating system (OS) vendor publishes a security patch to defeat a particular attack scheme, the hackers have figured out another way to defeat the software. Once viruses and the like appear on servers, an entire network of computers is susceptible to attack by those viruses.

[0003] In addition to malicious attacks in which the intent is to cause widespread system damage, networks are also prone to security breaches that enable data to be "stolen." For example, recent attacks have been made on various electronic storefront servers to steal credit card information and other user information. These types of attacks have lead to an escalating need for substantially improved security measures.

[0004] In view of the severity and frequency of the foregoing, a new direction has been proposed to replace today's security paradigm. A more proactive approach to security is presently being designed into the next generation of operating systems, which are referred to as trusted operating systems (TOS), secure operating systems (SOS), and secure and trusted operating systems (STOS). As stated in an *NSA Operating System Security Paper, NISSC, October 1998*, "Current security efforts

suffer from the flawed assumption that adequate security can be provided in applications with the existing security mechanisms of mainstream operating systems. In reality, the need for secure operating systems is growing in today's computing environment due to substantial increases in connectivity and data sharing. The threats posed by the modern computing environment cannot be addressed without secure operating systems. Any security effort which ignores this fact can only result in a 'fortress built upon sand'."

[0005] In contrast to today's scheme of security mechanisms layered over an unsecure core (e.g., a mainstream OS), the new approach begins with a trusted core that may only be accessed by users having appropriate security credentials. In this context, it is noted that users are not limited to humans, but rather also include programmatic entities such as software applications and the like. A chain of trust is maintained by the TOS or STOS to ensure that only trustworthy users may access secured portions of the OS, while other unsecure portions do not require the same level of authentication to access. The end result is that unqualified access is denied.

[0006] While efforts are being made to dramatically enhance operating system security, there are no commensurate endeavors that have been directed at system firmware. Although less susceptible to attacks than software, there are still situations under which system firmware may be attacked. In general, an attack may be direct (e.g., immediate deletion or corruption of critical BIOS components), or latent (e.g., a firmware Trojan to be activated at a future time). While operating systems attacks are generally somewhat recoverable, a firmware attack may be fatal. Consider, an errant operating system can be replaced, while errant firmware may totally disable a system or server.

[0007] Another aspect of firmware security relates to trusted processes. STOS's provide mechanisms to ensure all secure software processes are trusted. There are no equivalent mechanisms for ensuring firmware processes are trusted; as such,

firmware is deemed to be untrustworthy. For example, various firmware architectures support "hidden" firmware operations during operating system runtime. Depending on the particular architecture, these runtime firmware operations may have access to all of a system's resources, including the full memory address space.

- 5 Thus, malicious firmware may cause havoc to operating systems an/or applications by modifying (e.g., deleting, replacing) portions of memory that are thought to be secured by the TOS or STOS.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The foregoing aspects and many of the attendant advantages of this invention will become more readily appreciated as the same becomes better understood by reference to the following detailed description, when taken in conjunction with the accompanying drawings, wherein like reference numerals refer to like parts throughout the various views unless otherwise specified:

[0009] Figure 1 is a schematic diagram of an exemplary platform configuration via which embodiments of the invention may be implemented;

[0010] Figure 2 is a flowchart illustrating operations and logic performed during initialization of a system to securely store measurements of trusted firmware, according to one embodiment of the invention;

[0011] Figure 3 is a flowchart illustrating operations and logic performed during an operating system (OS)-runtime evaluation, wherein secure firmware processing is effectuated, according to one embodiment of the invention;

[0012] Figure 4 is a flowchart illustrating operations and logic performed in response to a system management interrupt (SMI) or a platform management interrupt (PMI) event, wherein a trusted firmware process may be effectuated, according to one embodiment of the invention;

[0013] Figure 5 is a schematic diagram illustrating a technique for storing a secret in a platform configuration register (PCR) having a trusted locality, according to one embodiment of the invention;

[0014] Figure 6a is a schematic diagram illustrating a technique for sealing a secret to a trusted platform module (TPM), wherein the secret is contained in a digest including one or more integrity metric measurements, at least one of which is stored in a PCR having a trusted locality, according to one embodiment of the invention;

[0015] Figure 6b is a schematic diagram illustrating a technique for sealing a secret to a TPM, wherein the secret is contained in a digest including one or more integrity metric measurements plus indicia identifying a locality that was asserted at the time the secret is sealed, wherein at least one of the measurements is stored in a PCR having a trusted locality, according to one embodiment of the invention; and

[0016] Figure 6c is a schematic diagram illustrating a technique for sealing a secret to a TPM, wherein the secret is contained in a digest including one or more integrity metric measurements plus indicia identifying a locality that was asserted at the time the secret is sealed, according to one embodiment of the invention.

10

DETAILED DESCRIPTION

[0017] Embodiments of systems and methods for supporting platform firmware as a trusted process are described herein. In the following description, numerous specific details are set forth to provide a thorough understanding of embodiments of the invention. One skilled in the relevant art will recognize, however, that the invention can be practiced without one or more of the specific details, or with other methods, components, materials, *etc.* In other instances, well-known structures, materials, or operations are not shown or described in detail to avoid obscuring aspects of the invention.

[0018] Reference throughout this specification to “one embodiment” or “an embodiment” means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, the appearances of the phrases “in one embodiment” or “in an embodiment” in various places throughout this specification are not necessarily all referring to the same embodiment. Furthermore, the particular features, structures, or characteristics may be combined in any suitable manner in one or more embodiments.

[0019] In accordance with aspects of the embodiments disclosed herein, trusted firmware processes are effectuated via trusted platform technology in combination with a secure authentication credential storage scheme. These security measures ensure the trustworthiness of firmware, while preventing attacks via execution of malicious firmware.

[0020] An exemplary computer architecture 100 suited for implementing embodiments of the invention described herein is shown in Figure 1. The architecture includes a processor 102 coupled, via a bus 104, to a memory controller hub (MCH) 106, commonly referred to as the “Northbridge” under well-known Intel® chipset schemes. MCH 106 is coupled via respective buses 108 and 110 to system

memory (*i.e.*, RAM) 112 and an advanced graphics port (AGP) 114. MCH 106 provides memory control functions, and includes a device protection mechanism 111 that enables access to protected memory pages 113 in system memory 112. MCH 106 is further coupled to an Input/Output (I/O) controller hub (ICH) 116 via a bus 118. The ICH, which is commonly referred to as the "Southbridge," provides a hardware interface to various I/O buses, ports and devices, depicted as items 120. These include a PCI bus, and IDE interface, a universal serial bus (USB), *etc.* ICH 116 is further coupled to a network port 122 via an I/O path 124. In turn, the network port 122 may be coupled to a network 125.

[0021] The architecture 100 of Figure 1 also includes a firmware hub 126 coupled to ICH 116 via a low pin count (LPC) bus 128, which is configured per Intel LPC Interface Specification Revision 1.0, September 29, 1997. The firmware hub includes various firmware components, including a BIOS boot block 130 and replaceable or modular firmware code 131. The architecture may include a non-volatile (NV) store 134 in which non-volatile data used by the firmware and/or other components may be stored.

[0022] In accordance with one aspect of the embodiments, a Trusted Computing Group (TCG) (<http://www.trustedcomputinggroup.org>) security scheme is implemented to store and retrieve security-related data. In the embodiment of Figure 1, a TCG token comprising a trusted platform module (TPM) is employed. Generally, TPM functionality may be embodied as a hardware device (most common) or via software. For example, Integrated circuits have been recently introduced to support TPM functionality, such as National Semiconductor's LPC-based TCG-compliant security controller. Such an integrated circuit is depicted as a TPM 136 in Figure 1.

[0023] TCG is an industry consortium concerned with platform and network security. The TCG main specification (Version 1.2, October, 2003 – hereinafter

referred to as the "version 1.2 specification") is a platform-independent industry specification that covers trust in computing platforms in general. The TCG main specification defines a trusted platform subsystem that employs cryptographic methods when establishing trust. The trusted platform may be embodied as a device or devices, or may be integrated into some existing platform component or components. The trusted platform enables an authentication agent to determine the state of a platform environment and seal data particular to that platform environment. Subsequently, authentication data (e.g., integrity metrics) stored in a TPM may be returned in response to an authentication challenge to authenticate the platform.

[0024] A "trusted measurement root" measures certain platform characteristics, logs the measurement data, and stores the final result in a TPM (which contains the root of trust for storing and reporting integrity metrics). When an integrity challenge is received, the trusted platform agent gathers the following information: the final results from the TPM, the log of the measurement data from the trusted platform measurement store, and TCG validation data that states the values that the measurements should produce in a platform configured in accordance with the configuration that existed at the time the integrity measurements were sealed. The operations of making an identity and enabling key-pair for the pre-boot environment enables TPM functionality to be employed for authentication purposes during and after pre-boot. Further details concerning the use of TPM 136 are discussed below.

[0025] TPM 136 provides several functions relating to security. These include an cryptographic co-processor 138, an HMAC (Hashing for Message Authentication Codes) engine 140, and SHA (security hash algorithm) -1 engine 142, an Opt-In component 144, non-volatile memory 146, a key generator 148, a random number generator (RNG) 150, an execution engine 152, volatile memory 156, and Platform Configuration Registers (PCRs) 156. Also provided in one TPM embodiment but not shown are an input/output component and a power detection component.

[0026] Generally, a TPM by itself provides a baseline level of security for storing and accessing trust-related data and authentication measures. Under TPM Main Specification, Part 1 Design Principles (October 2, 2003) a TPM is to be an independent device that is not susceptible to tampering or incorrect usage.

5 Accordingly, to further enhance this baseline security, an embodiment of the invention implements a hidden access mechanism that enables access to TPM 136 via special bus cycles invoked on low pin count (LPC) bus 128.

[0027] In order to provide for trust processes, there must be a mechanism to ensure that security-related data, such as secrets (*e.g.*, keys, certificates, *etc.*) may
10 not be accessed by unqualified users. In the context of TPM usage, users may comprise software, firmware or hardware; for convenience, all users are generally referred to as software. The TPM architecture addresses this aspect using a mechanism referred to as "locality." The concept of locality was introduced in the TPM Main Specification version 1.2. Locality allows for Dynamic RTMs in addition to
15 static RTMs. Locality also allows more than one simultaneous root of trust to exist per platform operational state.

[0028] The idea behind locality is that certain combinations of software and hardware are allowed more privileges than other combinations. This is enforced by a hierarchical access structure. For instance, the highest level of locality might be
20 cycles that only hardware could create. Since the hardware is virus-proof (at least in theory), it has a higher level of trust than cycles code generates. Under the version 1.2 specification, Locality 4 is the highest level defined. These cycles are generated by hardware. This would include things like TPM_PCR_Reset or the TPM_HASH_* LPC commands. The hardware guarantees that the TPM_PCR_RESET or HASH
25 operations occurred, and that the correct data was sent.

[0029] Next, consider the case of a platform that has both: 1) software based on either using no PCRs or a set of resettable PCRs (the OS based on the Static RTM

or Static OS); and 2) trusted software (*i.e.*, a trusted OS). In this case, there is a need to differentiate cycles from the OS's. Localities 1-3 may be used by the TOS for its transactions. The version 1.2 specification includes definitions for new resettable PCRs assigned to specific localities. Only trusted software should be able
5 to issue commands based on those PCRs. This software is locality 1-3 software.

[0030] With reference to the flowchart of Figure 2, in one embodiment a platform is initialized to support a trusted firmware process effectuated via locality in the following manner. The process starts in a block 200 in response to a restart event. A restart event implies the system has been restarted, such as via a cold boot or
10 system reset (*a.k.a.*, warm boot). In response, processor, chipset, and memory initialization operations are performed. These include initializing the platform memory and I/O complex, as depicted in a block 202.

[0031] Next, a determination is made to whether the platform has a TPM in a decision block 204. In order to enforce trustworthiness, there must be some
15 component(s) that are inherently deemed trustworthy. This component or components are used to derive the core root of trust measurement (CRTM). A CRTM is the basis for all subsequent trust measurements. The TPM is used to store these measurements.

[0032] If a TPM exists, the answer to decision block 204 is YES, and the logic
20 proceeds to a block 206. In this block, the TPM_STARTUP command is issued. A static CRTM is then created via a measurement of the platform's base firmware. The base firmware is a portion of firmware provided by the platform that cannot be partially replaced. The CRTM must be an immutable portion of the platform's initialization code that executes upon any platform reset. The trust in the platform is
25 based on the CRTM. The trust in all measurements is based on the integrity of this component. Thus, the static CRTM is derived from a measurement of "static" firmware.

[0033] Currently, in a PC, there are at least two types of static CTRM architectures. The first CTRM architecture measures the BIOS boot block 130 (a.k.a., firmware boot block or trusted boot block). In this architecture, the firmware is composed of the BIOS boot block and a POST (power-on self-test) BIOS. These
5 are individual components and each can be updated independent of the other. In this architecture, the BIOS boot block is the CTRM while the POST BIOS is not, but is a measured component of the Chain of Trust. If there are multiple execution points for the BIOS boot block, they must all be within the CTRM.

[0034] Under a second architecture, the CTRM is measured over the entire
10 BIOS. In this architecture, the BIOS is composed of a single atomic entity. The entire BIOS is updated, modified, or maintained as a single component, and cannot be partially replaced.

[0035] In some embodiments, a measured portion of the BIOS is used to measure the CTRM-applicable components, and then store the measurement in
15 PCR0 in the TPM, such as depicted by a static CTRM 158. In other embodiments, special processor microcode is used to perform the "dynamic" CTRM measurement and storage operations. Each of these techniques is performed in connection with inherent functionality provided by the TPM for such purposes.

[0036] In one embodiment, trusted firmware is assigned to locality 1. In one
20 embodiment, the processor must be in a secure execution mode (SEM) to enter locality 1. For IA-32 processors, the SEM is entered by issuing an "SENDER" instruction to processor 102. While in secure execution mode, all existing and potential future processor users are blocked from accessing the processor. SEM also temporarily blocks all interrupts (the interrupts are redirected for subsequent
25 handling after exiting SEM), including system management interrupts (SMIs), as depicted by an SMI redirection block 160.

[0037] In accordance with one aspect of SEM operation, processor 102 contains special instructions and microcode 162 to access certain devices coupled to LPC 142 via special bus cycle timing. These devices include TPM 136. This physical access scheme provides one level of security between data stored in
5 TPM 136 and attacks on platform architecture 100.

[0038] A second level of security is provided by storing integrity metric data in platform configuration registers 156. PCR's 156 are employed for securely storing data in a manner where certain authentication information must be provided to TPM 136 in order to access a given PCR. In particular, the processor controls the
10 locality, thus enabling certain PCR's from being accessed unless operating in the locality corresponding to those registers. Further security may be provided by sealing the locality in the measurement, as described below.

[0039] A PCR is a 160-bit storage location for discrete integrity measurements. All PCR registers are *shielded-locations* and are inside of the TPM. The decision of
15 whether a PCR contains a *standard measurement* or if the PCR is available for general use is deferred to the platform specific specification.

[0040] A large number of integrity metrics may be measured in a platform, and a particular integrity metric may change with time and a new value may need to be stored. It is difficult to authenticate the source of measurement of integrity metrics,
20 and as a result a new value of an integrity metric cannot be permitted to simply overwrite an existing value. (A rogue entity could erase an existing value that indicates subversion and replace it with a benign value.) Thus, if values of integrity metrics are individually stored, and updates of integrity metrics must be individually stored, it is difficult to place an upper bound on the size of memory that is required to
25 store integrity metrics.

[0041] A PCR is designed to hold an unlimited number of measurements in the register. It does this by using a cryptographic hash and hashing all updates to a PCR. The pseudo code for this is:

$$\text{PCR}_i \text{ New} = \text{HASH}(\text{PCR}_i \text{ Old value} \parallel \text{value to add})$$

5 Updates to a PCR register are sometimes referred to as “extending” the PCR, while the data measured to the PCR is sometimes called the “extend.” Since the resultant hash is dependant on existing data (the Old value), a chain of trust is formed that is incremented with each extend operation.

[0042] In one embodiment, PCR[0] provides the capability of being reset.
10 Accordingly, hash-extend operations may be performed in a manner that produces PCR[0] values that are independent of previously stored register values. This is advantageous with respect to being able to store integrity metrics corresponding to a given platform environment, and then subsequently compare integrity metrics corresponding to a current platform environment with the given platform
15 environment.

[0043] For example, in one embodiment of block 206, PCR[0] is reset, and a hash-extend is performed on the trusted BIOS boot block code (the core integrity measurement) using SHA-1 engine 142, with the result being stored in PCR[0]. In this context, the hash-extend operates on a reset register value (*i.e.*, 0), and so the
20 hash-extend simply reflects a hash of the trusted code. Thus, once loaded, the trusted BIOS boot block corresponds to one of the platform firmware environment components, while the hash of the component comprises an integrity metric corresponding to the platform environment, in this case the static CRTM. (It is noted that an integrity metric corresponding to a platform environment may reflect a single
25 environment component (*i.e.*, firmware/software component), or a combination of components used to form an environment that exists at the time the integrity metric is measured.)

[0044] Continuing with the flowchart of Figure 2, in a decision block 208 a determination is made to whether the processor supports dynamic CRTM measurements. With the extension of the locality functionality that was added to the TPM by the version 1.2 specification, both a static CRTM and a dynamic CRTM may exist. A dynamic CRTM is derived from the processor itself. It is sealed during platform manufacturing in PCR[17], which may only be accessed while in locality 4; a sealed dynamic CRTM 164 is shown in Figure 1.

[0045] If the answer to decision block 208 is YES, the logic proceeds to a block 210 in which an xMI_DESCRIPTOR data structure is created that indicates “where” the SMM or PMI-based startup code exists. Generally, SMM startup code relates to code that is executed by a “hidden” processor mode known as system management mode (SMM) on an IA-32 (Intel® Architecture, 32-bit) processor in response to a system management interrupt (SMI) event. Meanwhile, PMI (platform management interrupt) startup code relates to code that is executed in a somewhat analogous processor mode on an IA-64 processor, such as a member of the Intel® Itanium® family. For simplicity, both SMI and PMI signals are sometimes referred to herein as “xMI” signals. It will be understood that reference to an xMI signal means one of an SMI signal or a PMI signal.

[0046] As describe in further detail below, SMM- and PMI-based code comprises firmware that is loaded during the pre-boot into a reserved portion of memory. In block 210, the location at which the startup portion of this code is stored by the xMI_DESCRIPTOR. This portion of firmware is depicted as SMM or PMI startup code 132 in Figure 1. For SMM code, the startup code is also referred to as the SMBASE.

[0047] If there is no dynamic CRTM defined for the processor, the CRTM in the trusted building block (TBB) measures the SMM- or PMI-based startup code in a block 212. In these embodiments, the startup code concerns the portion of the

SMM- or PMI-based firmware that is fixed by the platform (e.g., at manufacture). There are firmware models that exist that enable third-party SMM and PMI code to be loaded for execution in response to corresponding SMI- and PMI-based events. This extensible portion of the SMM or PMI code, shown at 133, is not measured in
5 block 212.

[0048] Continuing with the initialization process, each of the logic flows for a NO result of decision block 204, or completion of either of blocks 210 or 212, lead to a decision block 214. In this decision block, a determination is made to whether the system employs an IA-32 processor. If it does, the SMBASE (*i.e.*, SMM- startup
10 code) is relocated from an address of 0x3000:00 (its default address in firmware storage) to one of the H- segment, T-segment, and/or the compatibility segment in a block 216. If the answer to decision block 214 is NO, the processor is an IA-64 processor, and the PMI entry point is registered with the IA-64 processor abstraction layer (PAL) in a block 218.

[0049] Following the completion of either of blocks 216 and 218 (as applicable), a
15 determination is made in a decision block 218 to whether there is dynamic CRTM defined for the processor. This is analogous to the determination made in decision block 208 discussed above. If the answer is YES, the processor microcode measures the PMI code or the SMM code referenced by the xMI_DESCRIPTOR
20 created in block 210 above, as depicted in a block 220. The initialization of the system is then completing and the operating system is booted in a block 222. If the answer to decision block 220 is NO, the operation of block 222 is skipped. After the completion of loading the OS, the system is in OS-runtime.

[0050] Operations performed to ensure a secure firmware-processing
25 environment during OS-runtime in accordance with one embodiment are shown in Figure 3. The process begins with an SMI or PMI event. These events correspond

to situations that cause the processor to switch from its current tasks to execute SMI or PMI-based code to service the event.

5 **[0051]** In response to the event, the system asserts locality 1 in a block 302. In practice, the processor first asserts locality 4, and then locality 1. In one embodiment, this is performed by issuing an SENTER command, causing execution control to pass to processor 102. The processor then executes microcode 162 to assert locality 4. The SMM or PMI startup code 132 then begins to execute. This startup code includes instructions to assert locality 1.

10 **[0052]** In a decision block 304 a determination is made to whether the event involves security processing. If it does, the logic flows to a block 306, wherein TPM ordinals are sent and results are returned. For example, the TPM ordinals may comprise a shared secret or certificate that is stored in either a PCR or a non-volatile storage means. This shared secret is extracted by the SMM or PMI code and provided to a challenger, who has requested authentication credentials. The results
15 correspond to information sent back to the client being challenged (*i.e.*, the platform hosting the TPM in which the credentials are sealed). After the completion of the operations in block 306, other SMM or PMI processing may be performed in a block 308.

20 **[0053]** Figure 4 shows a flowchart corresponding to one embodiment of a mechanism for performing trusted firmware processes during OS-runtime. It will be understood that the initialization operations of Figure 2 or similar operations will have been performed during the pre-boot prior to booting and operating system and entering the OS-runtime phase. The process starts with an SMI or PMI event in a block 400. In response, the current SMM or PMI startup code is measured in a
25 block 402. This measurement is analogous to the measurements performed in blocks 212 and 222 of Figure 2. Prior to this measurement, the SMM or PMI startup code is deemed "unqualified." This is because there is no way to verify whether the

current SMM or PMI startup code is trustworthy or not without some sort of qualification, such as the measurement performed in block 402

[0054] In a block 404, the system asserts locality 1. The original SMM or PMI code measurement is then retrieved from a locality 1 PCR in a block 406. A
5 determination to whether or not the SMM/PMI startup code measurements of blocks 402 and 406 match is made in a decision block 408. What this determination does, in effect, is verify that the current SMM or PMI startup code is identical to the startup code that was measured during the initialization operation of either block 212 or block 222. The startup code measured during the initialization is deemed
10 trustworthy. This mechanism authenticates the trustworthiness of the current SMM or PMI startup code.

[0055] If the answer to decision block 408 is YES, indicating that startup code is trustworthy, the processing is allowed to continue in locality 1, as depicted by a block 410. Thus, all locality 1 privileges are made available to any subsequent SMM
15 or PMI firmware processing performed in a block 412. If the SMM/PMI startup code does not match, the firmware is deemed untrustworthy. Thus, in a block 414, the locality 0 privilege level is asserted, and any subsequent SMM/PMI processing performed in block 412 may only operate in this locality, precluding the firmware from accessing locality 1 PCRs.

[0056] In the foregoing scheme, the operation of blocks 402, 406 and 408 may be
20 performed on a one-time basis (e.g., once per OS-runtime session), or every time an SMI or PMI event occurs. In accordance with one embodiment of the one-time basis, information is stored on the platform (e.g., at a known memory location, in NVRAM, etc.) indicating that the SMM/PMI startup code is deemed trustworthy. This
25 is depicted by an optional block 414.

Secure Processing Operations – Protected Security/Authentication Data

[0057] There are many advantageous to being able to operate in a locality above locality 0. A principle advantage concerns storing security and authentication data. These type of data are generally referred to as "secrets," and include such objects
5 as keys and authentication certificates. The embodiments described below employ both physical mechanisms and logical mechanisms to prevent unauthorized access to such secrets.

[0058] As discussed above, the asserted locality corresponds to a privileged access level within a hierarchy, wherein the higher the locality, the higher the
10 hierarchy level and thus privilege. Also, the hierarchy level reflect trustworthiness of the process – the processes that are deemed the most trustworthy are allowed access to all operations, while less trustworthy processes are denied access to certain operations, based on the locality asserted for the processes.

[0059] Under the most basic secret-protection scheme, secrets may be stored
15 directly in PCR's corresponding to a processes' asserted locality. For example, as shown in Figure 5, a secret 500 is stored to a PCR[20]. The following table defines the PCR usage for the various localities defined by the version 1.2 specification:

Locality Modifier	Entity in Control	PCRs that can be reset by the locality	PCR that can be extended by this locality
Any	Any Software	15	15
0	Platform or Static OS	None	0 – 14, 23
1	Dynamic OS	None	20
2	Dynamic OS	20, 21, 22	20, 21, 22
3	An auxiliary level of a trusted component	None	18, 19, 20
4	Hardware / Dynamic RTM	17, 18, 19, 20	17, 18

TABLE 1

[0060] The combination of the TPM and the processor effectuate a physical security enforcement mechanism in the embodiment of Figure 5. Each locality level "inherits" the privileges of all of the localities below it. Thus, the process for which the locality is asserted in Figure 5 must have a locality of 1 or higher.

[0061] Designers of secure distributed systems, when considering exchange of information between systems, must identify the endpoints of communication. The composition and makeup of the endpoint is as important to the overall security of the system as is the communications protocol. As specified by the TCG, endpoints are minimally comprised of asymmetric keys, key storage and processing that protects protocol data items.

[0062] Classic message exchange based on asymmetric cryptography suggests that messages intended for one and only one individual can be encrypted using a public key. Furthermore, the message can be protected from tampering by signing with the private key. Keys are communication endpoints and improperly managed keys can result in loss of security. Additionally, improperly configured endpoints

may also result in loss of security. The TPM aids in improving security by providing both key management and configuration management features (e.g., Protected Storage, Measurement and Reporting). These features can be combined to “seal” keys and platform configuration making endpoint definition stronger.

5 **[0063]** TCG defines four classes of protected message exchange; Binding, Signing, Sealed-Binding (a.k.a. Sealing) and Sealed-Signing. *Binding* is the traditional operation of encrypting a message using a public key. That is, the sender uses the public key of the intended recipient to encrypt the message. The message is only recoverable by decryption using the recipient’s private key. When the private
10 key is managed by the TPM as a non-migratable key, only the TPM that created the key may use it. Hence, a message encrypted with the public key, “bound” to a particular instance of a TPM.

[0064] *Sealing* takes binding one step further. Sealed messages are bound to a set of platform metrics specified by the message sender. Platform metrics specify
15 platform configuration state that must exist before decryption will be allowed. Sealing associates the encrypted message (actually the symmetric key used to encrypt the message) with a set of PCR register values and a non-migratable asymmetric key.

[0065] A sealed message is created by selecting a range of PCR register values
20 and asymmetrically encrypting the PCR values plus the symmetric key used to encrypt the message. The TPM with the asymmetric decryption key may only decrypt the symmetric key when the platform configuration matches the PCR register values specified by the sender. Sealing is a powerful feature of the TPM. It provides assurance that protected messages or secrets are only recoverable when
25 the platform is functioning in a very specific known configuration.

[0066] Each of the embodiments of Figures 6a, 6b, and 6c involve the use of sealing information to a TPM. The techniques used for these purposes may be used

for stand-alone purposes, such as guaranteeing process security, or may be combined with operations involving another system, such as an attestation performed via a message exchange.

[0067] Sealing is effectuated via the TPM_Seal command. The SEAL operation
5 allows software to explicitly state a future "trusted" configuration that the platform must be in for the secret (stored via the TPM_Seal command) to be revealed. The SEAL operation also implicitly includes the relevant platform configuration (PCR-values) when the SEAL operation was performed. The SEAL operation uses the tmpProof value to BIND a BLOB (Binary Large Object) to an individual TPM. To
10 retrieve the secret, and UNSEAL operation is performed. If the UNSEAL operation succeeds, proof of the platform configuration that was in effect when the SEAL operation was performed is returned to the caller, as well as the secret data.

[0068] In response to the TPM_Seal command, external data is concatenated with a value of integrity metric sequence and encrypted under a parent key. The
15 TPM_Unseal command may be subsequently used to decrypt the BLOB using the parent key and export the plaintext data if the current integrity metric sequence inside the TPM matches the value of integrity metric sequence inside the BLOB.

[0069] With reference to Figure 6a, an embodiment is shown that seals a secret 600 to TPM 156 using one or more integrity metrics. During pre-boot, various
20 platform integrity metrics are measured and stored in respective PCR's. In the illustrated embodiment, these include a CRTM 602, SMM/PMI startup code 132, and an optional integrity metric 604. The CRTM measurement is stored in PCR[0], while the SMM/PMI startup code measurement is stored in PCR[20]. The optional integrity metric 604 is stored in a PCR[M]. Each of PCR[0] and PCR[M] have a
25 locality requirement of 0. PCR[20] requires a locality of 1 or greater. Thus, in order to store an integrity metric in PCR[20], locality 1 or greater must be asserted.

[0070] The secret 600 is concatenated with the values in PCR[0], PCR[M] (if an optional integrity measurement was made), and PCR[20]. The resulting digest 606 is then sealed to TPM 156 using the TMP_Seal command.

[0071] As discussed above, in order to access the secret 600, the platform configuration metrics that currently exist must match corresponding integrity metrics stored in PCR[0], PCR[M] (if an optional integrity measurement was made), and PCR[20]. As a result, the process must be in locality 1 to access the value in PCR[20], and the integrity metrics must match. Accordingly, the SMM/PMI startup code and the CRTM measurement (e.g., BIOS boot block 130) must be the same as corresponding components that existed at the measurements were stored in PCR[0], PCR[M] (optional), and PCR[20]. If the integrity metrics match, the secret 600 can be returned by issuing the TMP_Unseal command (with the appropriate command parameters).

[0072] The embodiment of Figure 6b is analogous to that shown in Figure 6a, except for it goes one step further. While the embodiment of Figure 6a prevents locality 0 users from accessing secret 600 (no matter what the current integrity metrics are), it does not prevent users in localities 2 and 3 from accessing the secret. (It is noted that the locality 4 user, the processor, would have no motivation to access secret 600, although it would also not be precluded from accessing all of the PCR's for in which the trusted configuration integrity metric data are stored.)

[0073] The embodiment of Figure 6b addresses this potential problem by including locality information 608 in a digest 606A identifying that locality 1 was asserted when the digest was sealed to TPM 156. Thus, this embodiment requires locality 1 to be asserted while the secret 600 is attempted to be retrieved using TPM_Unseal. As a result, users in any locality other than locality 1 may not access secret 600 under any condition.

[0074] The embodiment of Figure 6c is analogous to that shown in Figure 6b, except in this instance all of the integrity metrics are stored in locality 0 PCR's. One disadvantage with this approach is that the integrity metric in PCR[M] may be extended by a locality 0 user, which would cause a denial of service. (This
5 deficiency may also exist with both of the embodiments of Figures 6a and 6b if an optional integrity metric is used and the value of PCR[M] is extended.)

[0075] It is noted that in each of the embodiments of Figures 6a-c, one or more of the integrity metrics may actually be used in generating the respective digests 600A, 600B, and 600C. It is further noted that optional integrity metrics may typically relate
10 to other firmware components, such as Option ROMs. In this case, PCR[2] is designated for storing Option ROM code measurements, while PCR[3] is designated for storing Option ROM configuration measurements. It should be unlikely that a non-software locality 0 user would attempt to extend either PCR[2] or PCR[3], unless motivated by malicious intent. Furthermore, a rogue software entity would need to
15 provide some mechanism for accessing the TPM in the first place, which again is not likely.

[0076] Many of the various operations performed during initialization and OS-runtime use of the embodiments described above are enabled via execution of firmware instructions. Thus, embodiments of this invention may be used as or to
20 support a software/firmware program or module executed upon some form of processing core (such as the computer processor) or otherwise implemented or realized upon or within a machine-readable medium. A machine-readable medium includes any mechanism for storing or transmitting information in a form readable by a machine (e.g., a computer). For example, a machine-readable medium can
25 include, for example, a read only memory (ROM); a random access memory (RAM); a magnetic disk storage media; an optical storage media; and a flash memory device, etc. In addition, a machine-readable medium can include propagated

signals such as electrical, optical, acoustical or other form of propagated signals (e.g., carrier waves, infrared signals, digital signals, etc.). For instance, all or a portion of the firmware instructions may be loaded from a network store.

5 **[0077]** The above description of illustrated embodiments of the invention, including what is described in the Abstract, is not intended to be exhaustive or to limit the invention to the precise forms disclosed. While specific embodiments of, and examples for, the invention are described herein for illustrative purposes, various equivalent modifications are possible within the scope of the invention, as those skilled in the relevant art will recognize.

10 **[0078]** These modifications can be made to the invention in light of the above detailed description. The terms used in the following claims should not be construed to limit the invention to the specific embodiments disclosed in the specification and the claims. Rather, the scope of the invention is to be determined entirely by the following claims, which are to be construed in accordance with established doctrines
15 of claim interpretation.